Model averaging for sparse seemingly unrelated regression using Bayesian networks among the errors

Supplementary Paper

Abdul Salam and Marco Grzegorczyk
Bernoulli Institute
Groningen University
E-mail: m.a.grzegorczyk@rug.nl

TABLE OF CONTENTS

S1 - Marginalisation over covariance matrix

Given the covariate matrix $\mathcal{D} \in \mathbb{R}^{N,S}$ with regression vector $\boldsymbol{\beta}_{\mathcal{D}}$, we build the corresponding design matrices $\mathbf{X}_t \in \mathbb{R}^{S,k}$ with Equation (10) from the main paper, and we define the error vectors $\mathbf{e}_t := \mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}_{\mathcal{D}}$. From Equation (8) of the main paper it then follows:

$$\mathbf{e}_t \sim \mathcal{N}_S(\mathbf{0}, \boldsymbol{\Sigma}) \qquad (t = 1, \ldots, T)$$

We introduce the short notation $\mathbf{E} := \{\mathbf{e}_1, \ldots, \mathbf{e}_T\}$. In terms of the precision matrix, $\mathbf{W} := \boldsymbol{\Sigma}^{-1}$, we then have the error likelihood:

$$\begin{aligned}
p(\mathbf{E}|\mathbf{W}) &= \prod_{t=1}^{T} p(\mathbf{e}_t|\mathbf{W}) \\
&= \prod_{t=1}^{T} (2\pi)^{-S/2} \cdot \det(\mathbf{W})^{1/2} \cdot \exp\left(-\frac{1}{2}\mathbf{e}_t^{\mathsf{T}}\mathbf{W}\mathbf{e}_t\right) \\
&= (2\pi)^{-\frac{S\cdot T}{2}} \cdot \det(\mathbf{W})^{\frac{T}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\mathbf{e}_t^{\mathsf{T}}\mathbf{W}\mathbf{e}_t\right) \\
&= (2\pi)^{-\frac{S\cdot T}{2}} \cdot \det(\mathbf{W})^{\frac{T}{2}} \cdot \exp\left(-\frac{1}{2}tr\left(\sum_{t=1}^{T}\mathbf{e}_t \cdot \mathbf{e}_t^{\mathsf{T}}\mathbf{W}\right)\right) \\
&= (2\pi)^{-\frac{S\cdot T}{2}} \cdot \det(\mathbf{W})^{\frac{T}{2}} \cdot \exp\left(-\frac{1}{2}tr\left(\mathbf{S}\mathbf{W}\right)\right)
\end{aligned}$$

where $\mathbf{S} := \sum_{t=1}^{T} \mathbf{e}_t \cdot \mathbf{e}_t^{\mathsf{T}}$.

On the precision matrix, $\mathbf{W}$, we impose a Wishart prior with scale matrix $\mathbf{V}$ and $\alpha > S + 1$ degrees of freedom, symbolically $\mathbf{W} \sim \mathcal{W}(\mathbf{V}, \alpha)$. The prior density is then given by:

$$p(\mathbf{W}) = \frac{\det(\mathbf{W})^{(\alpha-S-1)/2}}{2^{\alpha\cdot S/2} \cdot \det(\mathbf{V})^{\alpha/2} \cdot \Gamma_S(\frac{\alpha}{2})} \cdot \exp(-\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{W}))$$

where $tr(.)$ is the trace operator, and $\Gamma_q(.)$ is the multivariate Gamma function:

$$\Gamma_q(a) = \pi^{q(q-1)/4} \cdot \prod_{j=1}^{q} \Gamma\left(a + \frac{1-j}{2}\right)$$

This yields for the errors the marginal likelihood:

$$\begin{aligned}
p(\mathbf{E}) &= \int p(\mathbf{E}|\mathbf{W}) \cdot p(\mathbf{W})\, d\mathbf{W} \\
&= \frac{1}{c_1} \cdot \int \det(\mathbf{W})^{(\alpha+T-S-1)/2} \cdot \exp\left(-\frac{1}{2}tr\left((\mathbf{V}^{-1} + \mathbf{S})\mathbf{W}\right)\right)\, d\mathbf{W} \\
&= \frac{c_2}{c_1} \cdot \int \frac{1}{c_2} \cdot \det(\mathbf{W})^{(\alpha+T-S-1)/2} \cdot \exp\left(-\frac{1}{2}tr\left((\mathbf{V}^{-1} + \mathbf{S})\mathbf{W}\right)\right)\, d\mathbf{W} \\
&= \frac{c_2}{c_1} \cdot 1
\end{aligned}$$

where

$$\frac{1}{c_1} := (2\pi)^{-\frac{S\cdot T}{2}} \cdot \frac{1}{2^{\alpha\cdot S/2}\cdot \det(\mathbf{V})^{\alpha/2}\cdot \Gamma_S(\frac{\alpha}{2})}$$

and

$$c_2 := 2^{(\alpha+T)\cdot S/2}\cdot \det\left((\mathbf{V}^{-1}+\mathbf{S})^{-1}\right)^{(\alpha+T)/2}\cdot \Gamma_S(\frac{\alpha+T}{2})$$

Thus, we have:

$$\begin{aligned} p(\mathbf{E}) &= \frac{c_2}{c_1} \\ &= (2\pi)^{-\frac{S\cdot T}{2}} \cdot \frac{2^{(\alpha+T)\cdot S/2}\cdot \det\left((\mathbf{V}^{-1}+\mathbf{S})^{-1}\right)^{(\alpha+T)/2}\cdot \Gamma_S(\frac{\alpha+T}{2})}{2^{\alpha\cdot S/2}\cdot \det(\mathbf{V})^{\alpha/2}\cdot \Gamma_S(\frac{\alpha}{2})} \\ &= \pi^{-\frac{S\cdot T}{2}} \cdot \frac{\Gamma_S(\frac{\alpha+T}{2})}{\Gamma_S(\frac{\alpha}{2})}\cdot \det(\mathbf{V}^{-1})^{\alpha/2}\cdot \det\left(\mathbf{V}^{-1}+\mathbf{S}\right)^{-(\alpha+T)/2} \end{aligned}$$

where the inverse of the scale matrix, $\mathbf{V}^{-1}$, is the parametric matrix of the Wishart prior.

S2 - Conditional Gaussian distributions

Let the random vector $\mathbf{y} = (y_1, \ldots, y_S)^\mathsf{T} \in \mathbb{R}^S$ be multivariate Gaussian distributed, and consider the following partition into two random subvectors $\mathbf{y_a} \in \mathbb{R}^{S_a}$ and $\mathbf{y_b} \in \mathbb{R}^{S_b}$ with $S_a + S_b = S$:

$$\mathbf{y} = \begin{pmatrix}\mathbf{y_a}\\ \mathbf{y_b}\end{pmatrix} \sim \mathcal{N}_S\left(\begin{pmatrix}\boldsymbol{\mu_a}\\ \boldsymbol{\mu_b}\end{pmatrix},\ \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}} & \boldsymbol{\Sigma}_{\mathbf{a},\mathbf{b}}\\ \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{a}} & \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{b}}\end{pmatrix}\right)$$

We then have the marginal distributions $\mathbf{y_a} \sim \mathcal{N}_{S_a}(\boldsymbol{\mu_a}, \boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}})$ and $\mathbf{y_b} \sim \mathcal{N}_{S_b}(\boldsymbol{\mu_b}, \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{b}})$ and the conditional Gaussian distribution

$$\mathbf{y_b}|\mathbf{y_a} \sim \mathcal{N}_{S_b}\left(\boldsymbol{\mu_b} + \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{a}}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}}^{-1}(\mathbf{y_a} - \boldsymbol{\mu_a})\ ,\ \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{b}} - \boldsymbol{\Sigma}_{\mathbf{b},\mathbf{a}}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}}^{-1}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{b}}\right)$$

Accordingly, when partitioning $\mathbf{y} = (y_1, \ldots, y_S)^\mathsf{T}$ into the first $S-1$ elements, $\mathbf{y_a} = (y_1, \ldots, y_{S-1})^\mathsf{T}$, and the last element y_S:

$$\mathbf{y} = \begin{pmatrix}\mathbf{y_a}\\ y_S\end{pmatrix} \sim \mathcal{N}_S\left(\begin{pmatrix}\boldsymbol{\mu_a}\\ \mu_S\end{pmatrix},\ \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}} & \boldsymbol{\Sigma}_{\mathbf{a},S}\\ \boldsymbol{\Sigma}_{S,\mathbf{a}} & \boldsymbol{\Sigma}_{S,S}\end{pmatrix}\right)$$

we get:

$$y_S|(y_1, \ldots, y_{S-1}) \sim \mathcal{N}\left(\mu_S + \sum_{i=1}^{S-1}\tilde{\beta}_{i,S}(y_i - \mu_i)\ ,\ v_S^2\right)$$

where $\tilde{\beta}_{i,S} := \left(\boldsymbol{\Sigma}_{S,\mathbf{a}}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}}^{-1}\right)_{1,i}$ and $v_S^2 := \boldsymbol{\Sigma}_{S,S} - \boldsymbol{\Sigma}_{S,\mathbf{a}}\boldsymbol{\Sigma}_{\mathbf{a},\mathbf{a}}^{-1}\boldsymbol{\Sigma}_{\mathbf{a},S}$.
By successively applying this rule, the multivariate density of $\mathbf{y} = (y_1, \ldots, y_S)^\mathsf{T} \in \mathbb{R}^S$ can be factorized into a product of univariate conditional Gaussians

$$p(\mathbf{y}) = p(y_1)\cdot\prod_{s=2}^{S} p(y_s|y_1, \ldots, y_{s-1})$$

S3 - From graphs to covariance matrices

Consider a directed acyclic graph $\mathcal{G}$ for Gaussian distributed random variables $y_1, \ldots, y_S$, and assume - without loss of generality - that the variables are in topological order. This implies conditional Gaussian densities of the form (Geiger and Heckerman, 1994):

$$y_s|(y_1, \ldots, y_{s-1}) \sim \mathcal{N}\left(\mu_s + \sum_{i=1}^{s-1} \tilde{\beta}_{i,s}(y_i - \mu_i), v_s^2\right) \qquad (s = 1, \ldots, S)$$

where μ_s is the unconditional mean of y_s, v_s^2 is the conditional variance of y_s, and the regression coefficient $\tilde{\beta}_{i,s}$ $(i < s)$ reflects the strength of the relationship between y_i and y_s. There is an edge from y_i to y_s in $\mathcal{G}$ if and only if $\tilde{\beta}_{i,s} \neq 0$.

Shachter and Kenley (1989) show that this implies the joint multivariate Gaussian distribution:

$$\mathbf{y} := (y_1, \ldots, y_S)^\mathsf{T} \sim \mathcal{N}_S(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_S)^\mathsf{T} \in \mathbb{R}^S$ and covariance matrix $\boldsymbol{\Sigma} = \left\{\mathbf{W}_{(S)}\right\}^{-1}$ and that the precision matrix $\mathbf{W}_{(S)}$ can be computed via the recursion:

$$\mathbf{W}_{(1)} = \frac{1}{v_1^2} \quad \text{and} \quad \mathbf{W}_{(s)} = \begin{pmatrix} \mathbf{W}_{(s-1)} + \frac{\mathbf{b}_s \mathbf{b}_s^\mathsf{T}}{v_s^2} & -\frac{\mathbf{b}_s}{v_s^2} \\ -\frac{\mathbf{b}_s^\mathsf{T}}{v_s^2} & \frac{1}{v_s^2} \end{pmatrix} \qquad (s = 2, \ldots, S)$$

where

$$\mathbf{b}_s := (\tilde{\beta}_{1,s}, \ldots, \tilde{\beta}_{s-1,s})^\mathsf{T} \in \mathbb{R}^s \qquad (s = 2, \ldots, S)$$

are vectors of (partial) regression coefficients.

S4 - On sampling DAG-coherent precision matrices

In this section of the supplementary material we provide a more detailed description of the third step of our algorithm to sample a precision matrix $\mathbf{W}_\mathcal{G}$ that is coherent with a given DAG $\mathcal{G}$. For illustration purposes, we provide a concrete example at the end of this section.

In the second step (cf. Eq. (22) of the main paper) we have computed the univariate conditional Gaussians appearing in Eq. (17) of the main paper, namely:

$$e_s|(\mathbf{e}_{\tau_s}, \{\boldsymbol{\Sigma}_\mathcal{G}\}_{\{e_s,\tau_s\},\{e_s,\tau_s\}}) \sim \mathcal{N}(\mu_{e_s|\tau_s},\ \sigma^2_{e_s|\tau_s}) \qquad (s = 1, \ldots, S) \tag{1}$$

with conditional mean and conditional variance

$$\begin{aligned} \mu_{e_s|\tau_s} &:= \boldsymbol{\beta}_{e_s|\tau_s} \mathbf{e}_{\tau_s} \\ \sigma^2_{e_s|\tau_s} &:= \boldsymbol{\Sigma}_{e_s,e_s} - \boldsymbol{\Sigma}_{e_s,\tau_s} \left\{\boldsymbol{\Sigma}_{\tau_s,\tau_s}\right\}^{-1} \boldsymbol{\Sigma}_{\tau_s,e_s} \end{aligned}$$

and the vector of (partial) regression coefficients: $\boldsymbol{\beta}_{e_s|\tau_s} := \boldsymbol{\Sigma}_{e_s,\tau_s} \left\{\boldsymbol{\Sigma}_{\tau_s,\tau_s}\right\}^{-1}$.

That is, we have a regression coefficient $\beta_{i,s}$ for each $e_i \in \tau_s$. If e_i is the j-th element of τ_s, we define $\beta_{i,s}$ to be the j-th element of $\boldsymbol{\beta}_{e_s|\tau_s}$, and for all $e_i \notin \tau_s$ we set: $\beta_{i,s} := 0$.

We can then re-write Eq. (1) as:

$$\begin{aligned}
e_s|(\mathbf{e}_{\tau_s}, \{\boldsymbol{\Sigma}_{\mathcal{G}}\}_{\{e_s,\tau_s\},\{e_s,\tau_s\}}) &\sim \mathcal{N}(\mu_{e_s|\tau_s},\ \sigma^2_{e_s|\tau_s}) \\
&\sim \mathcal{N}(\boldsymbol{\beta}_{e_s|\tau_s}\mathbf{e}_{\tau_s},\ \sigma^2_{e_s|\tau_s}) && \text{(recall } \mu_{e_s|\tau_s} := \boldsymbol{\beta}_{e_s|\tau_s}\mathbf{e}_{\tau_s}\text{)} \\
&\sim \mathcal{N}\left(\sum_{i\in\tau_s}\beta_{i,s}e_i\ ,\ \sigma^2_{e_s|\tau_s}\right) && \text{(recall the definition of } \beta_{i,s}\text{)} \\
&\sim \mathcal{N}\left(\sum_{i=1}^{S}\beta_{i,s}e_i\ ,\ \sigma^2_{e_s|\tau_s}\right) && \text{(recall } \beta_{i,s} := 0 \text{ if } i \notin \tau_s\text{)} \\
&\sim \mathcal{N}\left(\mu_s + \sum_{i=1}^{S}\beta_{i,s}(e_i - \mu_i)\ ,\ \sigma^2_{e_s|\tau_s}\right) && \text{(with } \mu_i := 0 \text{ for all } i\text{)} \\
&\sim \mathcal{N}\left(\mu_s + \sum_{i=1}^{S}\beta_{i,s}(e_i - \mu_i)\ ,\ \nu^2_s\right) && \text{(with } \nu^2_s := \sigma^2_{e_s|\tau_s}\text{)}
\end{aligned}$$

The error DAG $\mathcal{G}$ implies a topological order on the errors.[1] Given the topological order $e_{\sigma(1)}, \ldots, e_{\sigma(S)}$, where $\sigma(.)$ is a permutation, we have for the parent sets

$$\tau_{\sigma(s)} \subset \{e_{\sigma(1)}, \ldots, e_{\sigma(s-1)}\} \tag{2}$$

We can re-write the above using the topological order $e_{\sigma(1)}, \ldots, e_{\sigma(S)}$:

$$e_{\sigma(s)}|(\mathbf{e}_{\tau_{\sigma(s)}}, \{\boldsymbol{\Sigma}_{\mathcal{G}}\}_{\{e_{\sigma(s)},\tau_{\sigma(s)}\},\{e_{\sigma(s)},\tau_{\sigma(s)}\}}) \sim \mathcal{N}\left(\mu_{\sigma(s)} + \sum_{i=1}^{S}\beta_{\sigma(i),\sigma(s)}(e_{\sigma(i)} - \mu_{\sigma(i)})\ ,\ v^2_{\sigma(s)}\right)$$

Because of the relationship: $\beta_{\sigma(i),\sigma(s)} = 0$ if $e_{\sigma(i)} \notin \tau_{\sigma(s)}$, it follows from Eq. (2):

$$e_{\sigma(s)}|(\mathbf{e}_{\tau_{\sigma(s)}}, \{\boldsymbol{\Sigma}_{\mathcal{G}}\}_{\{e_{\sigma(s)},\tau_{\sigma(s)}\},\{e_{\sigma(s)},\tau_{\sigma(s)}\}}) \sim \mathcal{N}\left(\mu_{\sigma(s)} + \sum_{i=1}^{s-1}\beta_{\sigma(i),\sigma(s)}(e_{\sigma(i)} - \mu_{\sigma(i)})\ ,\ v^2_{\sigma(s)}\right)$$

In the latter univariate conditional Gaussian distributions we recognize the structural form that is required for applying the recursion from Shachter and Kenley (1989), which allows the joint distribution of $\mathbf{e}_\sigma =: (e_{\sigma(1)}, \ldots, e_{\sigma(S)})^\mathsf{T}$ to be computed (see Supplement **S3**). The joint distribution is a multivariate Gaussian:

$$\mathbf{e}_\sigma \sim \mathcal{N}_S\left(\boldsymbol{\mu}_\sigma, \boldsymbol{\Sigma}_{\mathcal{G},\sigma}\right) = \mathcal{N}_S\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{G},\sigma}\right)$$

By rearranging the rows and columns of $\boldsymbol{\Sigma}_{\mathcal{G},\sigma}$ via the inverse permutation $\sigma^{-1}(.)$, so that the s-th row and column correspond to e_s again, we obtain the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{G}}$ that is coherent with the DAG $\mathcal{G}$, and

$$\mathbf{e} \sim \mathcal{N}_S\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{G}}\right)$$

[1] In $e_{\sigma(1)}, \ldots, e_{\sigma(S)}$ each error e_s succeeds its parent nodes in τ_s, so that all edges of $\mathcal{G}$ 'point from left to right'.

ILLUSTRATIVE EXAMPLE

Assume we have only $S = 3$ nodes (errors) e_1, e_2 and e_3 and the DAG $\mathcal{G}$:

$$\mathcal{G}: \quad e_1 \rightarrow e_2 \rightarrow e_3$$

This means we have the three parent sets:

$$\tau_1 = \{\ \}, \quad \tau_2 = \{e_1\}, \quad \tau_3 = \{e_2\}$$

To sample a precision matrix that is coherent with $\mathcal{G}$, symbolically $\mathbf{W}_{\mathcal{G}}$, we proceed as follows:

- First we sample a precision matrix of a full graph $\mathcal{G}_F$, symbolically $\mathbf{W}$.
  $\mathbf{W}$ is a prior (or posterior) sample from a 3-dimensional Wishart distribution. With probability 1 this $\mathbf{W}$ does not impose any conditional independences among the three nodes.

- We now exploit one of the key results from Geiger and Heckerman (2002), namely that the sampled $\mathbf{W}$ for a full graph $\mathcal{G}_F$ specifies the parameters of all possible DAGs.[2] That is, from $\mathbf{W}$ we can extract $\mathbf{W}_{\mathcal{G}}$ for any DAG $\mathcal{G}$.

- $\mathbf{W}$ is the precision matrix of a 3-dimensional Gaussian. When the mean vector is $\boldsymbol{\mu} = \mathbf{0}$, we have

  $$\mathbf{e} := (e_1, e_2, e_3)^{\mathsf{T}} \sim \mathcal{N}_3(\mathbf{0}, \mathbf{W}^{-1})$$

  We can factorize the joint Gaussian density into a product of three conditional Gaussian densities. For example:

  $$p(e_1, e_2, e_3|\mathcal{G}_F) = p(e_1) \cdot p(e_2|e_1) \cdot p(e_3|e_1, e_2)$$

  The conditional distributions are univariate Gaussians of the form:

  $$\begin{aligned} e_1 &\sim \mathcal{N}(0,\ \sigma_1^2) && (I) \\ e_2|e_1 &\sim \mathcal{N}(\beta_{1,2}e_1,\ \sigma_2^2) && (II) \\ e_3|(e_1, e_2) &\sim \mathcal{N}(\beta_{1,3}e_1 + \beta_{2,3}e_2,\ \sigma_3^2) && (III) \end{aligned}$$

  In Supplement **S2** we review how the parameters $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{2,3}$, σ_1^2, σ_2^2, and σ_3^2 of these conditional Gaussians can be computed from the parameters $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$ of the joint distribution.

- That is, we can transform the parameter space:

  $$\{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}\} \rightarrow \{\beta_{1,2}, \beta_{1,3}, \beta_{2,3}, \sigma_1^2, \sigma_2^2, \sigma_3^2\}$$

  The recursion from Shachter and Kenley (1989) allows us to transform back:

  $$\{\beta_{1,2}, \beta_{1,3}, \beta_{2,3}, \sigma_1^2, \sigma_2^2, \sigma_3^2\} \rightarrow \{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}\}$$

  so that we have a one-to-one mapping between the two parameter spaces:

  $$\{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}\} \leftrightarrow \{\beta_{1,2}, \beta_{1,3}, \beta_{2,3}, \sigma_1^2, \sigma_2^2, \sigma_3^2\}$$

  In Supplement **S3** we review the recursion from Shachter and Kenley (1989).

[2] This is only true under a set of assumptions. The Wishart distribution in combination with the Gaussian likelihood fulfills these assumptions; see Geiger and Heckerman (2002) for the details.

- But we can as well factorize the joint density in any other node order. For example, for the order e_2, e_3, e_1 we get for the full graph $\mathcal{G}_F$ the factorization:

$$p(e_1, e_2, e_3|\mathcal{G}_F) = p(e_2) \cdot p(e_3|e_2) \cdot p(e_1|e_2, e_3)$$

and the conditional distributions are univariate Gaussians again:

$$\begin{aligned} e_2 &\sim \mathcal{N}(0,\ \theta_2^2) && (I*) \\ e_3|e_2 &\sim \mathcal{N}(w_{2,3}e_2,\ \theta_3^2) && (II*) \\ e_1|(e_2, e_3) &\sim \mathcal{N}(w_{2,1}e_2 + w_{3,1}e_3,\ \theta_1^2) && (III*) \end{aligned}$$

whose parameters can as well be computed from the parameters $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$ of the joint multivariate Gaussian.

- For our given DAG

$$\mathcal{G}: \quad e_1 \to e_2 \to e_3$$

we need that conditional on e_2 the nodes e_3 and e_1 are independent.
That is, we need the factorization:

$$p(e_1, e_2, e_3|\mathcal{G}) = \prod_{s=1}^{3} p(e_s|\tau_s) = p(e_1) \cdot p(e_2|e_1) \cdot p(e_3|e_2)$$

Geiger and Heckerman (2002) show that the appearing conditional distributions must be identical to those computed earlier for the full graph.[3] That is, for the DAG $\mathcal{G}$ we must have the following composition of conditional Gaussian distributions:

$$\begin{aligned} e_1 &\sim \mathcal{N}(0,\ \sigma_1^2) && (I) \\ e_2|e_1 &\sim \mathcal{N}(\beta_{1,2}e_1,\ \sigma_2^2) && (II) \\ e_3|e_2 &\sim \mathcal{N}(w_{2,3}e_2,\ \theta_3^2) && (II*) \end{aligned}$$

Loosely speaking, the local conditional distributions are independent of the overall graph structure. Here for our concrete example this means that the conditional distribution (II*) of $e_3|e_2$ must be the same; independently of whether it is part of the full graph $\mathcal{G}_F$ or part of the DAG $\mathcal{G}$.

- To get the precision matrix that is coherent with $\mathcal{G}$, symbolically $\mathbf{W}_{\mathcal{G}}$, we think of the above system of conditional Gaussians as a special case of a factorization for the full graph $\mathcal{G}_F$ with some zero parameter(s):

$$\begin{aligned} e_1 &\sim \mathcal{N}(0,\ \sigma_1^2) && (I) \\ e_2|e_1 &\sim \mathcal{N}(\beta_{1,2}e_1,\ \sigma_2^2) && (II) \\ e_3|(e_2, e_1) &\sim \mathcal{N}(w_{2,3}e_2 + w_{1,3}e_1,\ \theta_3^2) && (II*) \end{aligned}$$

where in our example: $w_{1,3} := 0$.

[3] Also this is only true under a set of assumptions. The Wishart distribution and the Gaussian likelihood fulfill these assumptions; see Geiger and Heckerman (2002) for the mathematical details.

- We can apply the recursion from Shachter and Kenley (1989) to back transform these full conditional Gaussians (for node order e_1, e_2, e_3) into a precision matrix that is coherent with $\mathcal{G}$:

$$\{\beta_{1,2}, w_{1,3} = 0, w_{2,3}, \sigma_1^2, \sigma_2^2, \theta_3^2\} \leftrightarrow \{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}_\mathcal{G}\}$$

  We note that $\mathbf{W}_\mathcal{G}$ will be coherent with $\mathcal{G}$, since we have $w_{1,3} := 0$, so that

$$p(e_3|e_1, e_2) = p(e_3|e_2)$$

Prior probability of a precision matrix that is coherent with a DAG:

What does this mean for the prior probability $p(\mathbf{W}_\mathcal{G})$ of this particular $\mathbf{W}_\mathcal{G}$?

There is **no** explicit prior of a well-known form, from which $\mathbf{W}_\mathcal{G}$ could be sampled.

$\mathbf{W}_\mathcal{G}$ is sampled by sampling any $\mathbf{W}$ for a full graph $\mathcal{G}_F$ that fulfills

- It transforms into

$$\{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}\} \rightarrow \{\beta_{1,2}, \tilde{\beta}_{1,3}, \tilde{\beta}_{2,3}, \sigma_1^2, \sigma_2^2, \tilde{\sigma}_3^2\} \text{ for order } e_1, e_2, e_3$$

- and it transforms into

$$\{\boldsymbol{\mu} = \mathbf{0}, \mathbf{W}\} \rightarrow \{w_{2,3}, \tilde{w}_{2,1}, \tilde{w}_{3,1}, \tilde{\theta}_2^2, \theta_3^2, \tilde{\theta}_1^2\} \text{ for order } e_2, e_3, e_1$$

where the four regression parameters $\tilde{\beta}_{1,3}$,$\tilde{\beta}_{2,3}$,$\tilde{w}_{2,1}$, $\tilde{w}_{3,1}$ and the three variance parameters $\tilde{\sigma}_3^2$, $\tilde{\theta}_2^2 > 0$, $\tilde{\theta}_1^2 > 0$ can be arbitrary values, since they do not appear in the composition of the conditional Gaussians required for $\mathcal{G}$. The precision matrix that is coherent with $\mathcal{G}$, symbolically $\mathbf{W}_\mathcal{G}$, is fully specified by the remaining five parameters $\beta_{1,2}$, σ_1^2, σ_2^2, $w_{2,3}$, and θ_3^2.

S5 - Full conditional of unrestricted covariance matrix

For a given covariate matrix $\mathcal{D} \in \mathbb{R}^{N,S}$ with regression vector $\boldsymbol{\beta}_{\mathcal{D}}$, we defined: $\mathbf{e}_t := \mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta}_{\mathcal{D}}$ and $\mathbf{E} := \{\mathbf{e}_1, \ldots, \mathbf{e}_T\}$. For the posterior density of $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ we then have:

$$p(\mathbf{W}|\mathbf{E}) \propto p(\mathbf{E}|\mathbf{W}) \cdot p(\mathbf{W})$$

where $p(\mathbf{E}|\mathbf{W})$ is the error likelihood (see Supplement **S1**):

$$p(\mathbf{E}|\mathbf{W}) \quad = \quad (2\pi)^{-\frac{S \cdot T}{2}} \cdot \det(\mathbf{W})^{\frac{T}{2}} \cdot \exp\left(-\frac{1}{2} tr\left(\mathbf{S}\mathbf{W}\right)\right)$$

and $p(\mathbf{W})$ is the density of the Wishart prior (see Supplement **S1**):

$$p(\mathbf{W}) = \frac{\det(\mathbf{W})^{(\alpha-S-1)/2}}{2^{\alpha \cdot S/2} \cdot \det(\mathbf{V})^{\alpha/2} \cdot \Gamma_S(\frac{\alpha}{2})} \cdot \exp(-\frac{1}{2} tr(\mathbf{V}^{-1}\mathbf{W}))$$

We get for the posterior density:

$$\begin{aligned} p(\mathbf{W}|\mathbf{E}) &\propto p(\mathbf{E}|\mathbf{W}) \cdot p(\mathbf{W}|\mathbf{V}, \alpha) \\ &\propto \det(\mathbf{W})^{\frac{T}{2}} \cdot \exp\left(-\frac{1}{2} tr(\mathbf{S}\mathbf{W})\right) \cdot \det(\mathbf{W})^{(\alpha-S-1)/2} \cdot \exp\left(-\frac{1}{2} tr(\mathbf{V}^{-1}\mathbf{W})\right) \\ &\propto \det(\mathbf{W})^{(\alpha+T-S-1)/2} \cdot \exp\left(-\frac{1}{2} tr((\mathbf{S} + \mathbf{V}^{-1})\mathbf{W})\right) \end{aligned}$$

From the shape of the posterior density we conclude

$$\mathbf{W}|\mathbf{E} \sim \mathcal{W}\left((\mathbf{S} + \mathbf{V}^{-1})^{-1}, \alpha + T\right)$$

where $(\mathbf{S} + \mathbf{V}^{-1})^{-1}$ is the scale matrix of the posterior Wishart distribution.

S6 - Rules for Gaussian integrals

The following two rules have been taken from Section 2.3.3 in Bishop (1995).
Let $\mathbf{\Sigma}_1 \in \mathbb{R}^{S,S}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{k,k}$ be (positive definite) covariance matrices, let $\boldsymbol{\mu} \in \mathbb{R}^k$ denote a vector, and let $\mathbf{X} \in \mathbb{R}^{S,k}$ be a design matrix. For a Gaussian (regression) likelihood with a Gaussian prior on the regression parameter vector $\boldsymbol{\beta} \in \mathbb{R}^k$:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta} &\sim \mathcal{N}_S(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}_1) \\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\mu}, \mathbf{\Sigma}_2) \end{aligned}$$

we have the posterior distribution:

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}(\mathbf{\Sigma}_\dagger(\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_1^{-1}\mathbf{y} + \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu})\ ,\ \mathbf{\Sigma}_\dagger) \ \text{ where } \ \mathbf{\Sigma}_\dagger = (\mathbf{\Sigma}_2^{-1} + \mathbf{X}^\mathsf{T}\mathbf{\Sigma}_1^{-1}\mathbf{X})^{-1}$$

and the marginal distribution

$$\mathbf{y} \sim \mathcal{N}_S(\mathbf{X}\boldsymbol{\mu},\ \mathbf{\Sigma}_1 + \mathbf{X}\mathbf{\Sigma}_2\mathbf{X}^\mathsf{T})$$

S7 - Matrix lemma

Consider four matrices $\mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{U} \in \mathbb{R}^{n,k}$, $\mathbf{C} \in \mathbb{R}^{k,k}$, and $\mathbf{V} \in \mathbb{R}^{k,n}$. When the appearing matrix inverses exits, the **matrix inversion lemma** ('Woodbury's matrix identity') states that:

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

And the **matrix determinant lemma** states that:

$$\det(\mathbf{A} + \mathbf{UCV}) = \det(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}) \cdot \det(\mathbf{C}) \cdot \det(\mathbf{A})$$

S8 - Marginalisation over regression vector

Given the error DAG $\mathcal{G}$ with covariance matrix $\mathbf{\Sigma}_{\mathcal{G}}$ and the covariate matrix $\mathcal{D}$, we get (see Supplement **S6**) the marginal distribution (marginalised over $\boldsymbol{\beta}_{\mathcal{D}}$):

$$\mathbf{y}|(\mathbf{\Sigma}_{\mathcal{G}}, \mathcal{D}) \;\; \sim \;\; \mathcal{N}_{T\cdot S}(\mathbf{X}\boldsymbol{\mu}_0,\; \mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^{\mathsf{T}})$$

This implies the density:

$$p(\mathbf{y}|\mathbf{\Sigma}_{\mathcal{G}}, \mathcal{D}) = (2\pi)^{-\frac{T\cdot S}{2}} \cdot \det(\tilde{\mathbf{\Sigma}})^{-\frac{1}{2}} \cdot \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0)\}$$

where

$$\tilde{\mathbf{\Sigma}} := \mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^{\mathsf{T}} \;\; \in \mathbb{R}^{T\cdot S, T\cdot S}$$

Computation of the determinant
For the determinant $\det(\tilde{\mathbf{\Sigma}})$ the matrix determinant lemma (see Supplement **S7**) can be applied:

$$\begin{aligned}
\det(\tilde{\mathbf{\Sigma}}) &= \det\left(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^{\mathsf{T}}\right) \\
&= \det\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}})^{-1}\mathbf{X}\right) \cdot \det\left(\mathbf{\Sigma}_0\right) \cdot \det\left(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}}\right) \\
&= \det\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I}^{-1} \otimes \mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right) \cdot \det\left(\mathbf{\Sigma}_0\right) \cdot \det(\mathbf{I})^S \cdot \det(\mathbf{\Sigma}_{\mathcal{G}})^T \\
&= \det\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right) \cdot \det\left(\mathbf{\Sigma}_0\right) \cdot \det(\mathbf{\Sigma}_{\mathcal{G}})^T
\end{aligned}$$

Note that T denotes the number of observations (and not the matrix transpose: T).
We have:

$$\begin{aligned}
\mathbf{X}^{\mathsf{T}}(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X} &= (\mathbf{X}_1^{\mathsf{T}}, \ldots, \mathbf{X}_T^{\mathsf{T}}) \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{G}}^{-1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{\Sigma}_{\mathcal{G}}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} \\
&= \sum_{t=1}^{T} \mathbf{X}_t^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{G}}^{-1} \mathbf{X}_t
\end{aligned}$$

so that

$$\det(\tilde{\mathbf{\Sigma}}) \;\; = \;\; \det\left(\mathbf{\Sigma}_0^{-1} + \sum_{t=1}^{T} \mathbf{X}_t^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{G}}^{-1} \mathbf{X}_t\right) \cdot \det\left(\mathbf{\Sigma}_0\right) \cdot \det(\mathbf{\Sigma}_{\mathcal{G}})^T$$

Computation of the quadratic form:
We also have to compute the quadratic form:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0)$$

with $\tilde{\mathbf{\Sigma}} := \mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^{\mathsf{T}} \;\; \in \mathbb{R}^{T\cdot S, T\cdot S}$

Recalling the matrix inversion lemma from Supplement **S7**

$$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

we get:

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}^{-1} &= \left(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^{\mathsf{T}}\right)^{-1} \\
&= (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}})^{-1} - (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}})^{-1}\mathbf{X}\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}})^{-1} \\
&= (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1}) - (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})
\end{aligned}$$

Henceforth, we can decompose:

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0) = (\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{A}}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0) - (\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{B}}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0)$$

where

$$\begin{aligned}
\tilde{\mathbf{A}} &:= (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1}) \\
\tilde{\mathbf{B}} &:= (\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})
\end{aligned}$$

Furthermore, we define:

$$\tilde{\mathbf{C}} = \left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right)^{-1} = \left(\mathbf{\Sigma}_0^{-1} + \sum_{t=1}^{T}\mathbf{X}_t^{\mathsf{T}}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\right)^{-1} \in \mathbb{R}^{k,k}$$

and we decompose the vector $\tilde{\mathbf{y}} := (\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0) \in \mathbb{R}^{T\cdot S}$ with

$$\mathbf{y} = (\mathbf{y}_1^{\mathsf{T}},\ldots,\mathbf{y}_T^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{T\cdot S}, \quad \mathbf{X} = (\mathbf{X}_1^{\mathsf{T}},\ldots,\mathbf{X}_T^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{T\cdot S,k} \text{ and } \boldsymbol{\mu}_0 \in \mathbb{R}^k$$

into T subvectors

$$\tilde{\mathbf{y}} = \begin{pmatrix}\tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_T\end{pmatrix} = \begin{pmatrix}\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\mu}_0 \\ \vdots \\ \mathbf{y}_T - \mathbf{X}_T\boldsymbol{\mu}_0\end{pmatrix}$$

Then we have:

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0)^{\mathsf{T}}\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_0) = \tilde{\mathbf{y}}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\tilde{\mathbf{y}} - \tilde{\mathbf{y}}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\tilde{\mathbf{C}}\mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\tilde{\mathbf{y}}$$

And it follows from

$$\begin{aligned}
\tilde{\mathbf{y}}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X} &= (\tilde{\mathbf{y}}_1^{\mathsf{T}}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_1,\ldots,\tilde{\mathbf{y}}_T^{\mathsf{T}}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_T) \in \mathbb{R}^{1,k} \\
\mathbf{X}^{\mathsf{T}}(\mathbf{I}\otimes\mathbf{\Sigma}_{\mathcal{G}}^{-1})\tilde{\mathbf{y}} &= (\mathbf{X}_1^{\mathsf{T}}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\tilde{\mathbf{y}}_1,\ldots,\mathbf{X}_T^{\mathsf{T}}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\tilde{\mathbf{y}}_T) \in \mathbb{R}^{k,1}
\end{aligned}$$

that

$$
\begin{aligned}
(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0)^\mathsf{T}\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_0) &= \sum_{t=1}^{T} \tilde{\mathbf{y}}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\tilde{\mathbf{y}}_t \; - \; \sum_{t=1}^{T} \tilde{\mathbf{y}}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\tilde{\mathbf{C}}\mathbf{X}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\tilde{\mathbf{y}}_t \\
&= \sum_{t=1}^{T} \tilde{\mathbf{y}}_t^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{G}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\tilde{\mathbf{C}}\mathbf{X}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\right)\tilde{\mathbf{y}}_t \\
&= \sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\mu}_0)^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{G}}^{-1} - \boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\tilde{\mathbf{C}}\mathbf{X}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\right)(\mathbf{y}_t - \mathbf{X}_t\boldsymbol{\mu}_0)
\end{aligned}
$$

where

$$
\tilde{\mathbf{C}} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^\mathsf{T}(\mathbf{I} \otimes \boldsymbol{\Sigma}_{\mathcal{G}}^{-1})\mathbf{X}\right)^{-1} \;=\; \left(\boldsymbol{\Sigma}_0^{-1} + \sum_{t=1}^{T} \mathbf{X}_t^\mathsf{T}\boldsymbol{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\right)^{-1} \in \mathbb{R}^{k,k}
$$

S9 - Full conditional of regression vector

Given the error DAG $\mathcal{G}$ with covariance matrix $\mathbf{\Sigma}_{\mathcal{G}}$ and the covariate matrix $\mathcal{D}$, we have for the full conditional distribution of $\boldsymbol{\beta}_{\mathcal{D}}$:

$$p(\boldsymbol{\beta}_{\mathcal{D}}|\mathbf{\Sigma}_{\mathcal{G}}, \mathcal{D}, \mathbf{y}) \quad \propto \quad p(\mathbf{y}|\mathbf{\Sigma}_{\mathcal{G}}, \boldsymbol{\beta}_{\mathcal{D}}) \cdot p(\boldsymbol{\beta}_{\mathcal{D}})$$

where

$$\begin{aligned} \mathbf{y}|(\mathbf{\Sigma}_{\mathcal{G}}, \boldsymbol{\beta}_{\mathcal{D}}) &\sim \mathcal{N}_{T \cdot S}(\mathbf{X}\boldsymbol{\beta}_{\mathcal{D}}, \mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}}) \\ \boldsymbol{\beta}_{\mathcal{D}} &\sim \mathcal{N}_k(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0) \end{aligned}$$

A standard rule for Gaussian integrals (see Supplement **S6**) implies the full conditional distribution:

$$\boldsymbol{\beta}_{\mathcal{D}}|(\mathbf{\Sigma}_{\mathcal{G}}, \mathcal{D}, \mathbf{y}) \sim \mathcal{N}_k\left(\mathbf{\Sigma}^\star\left(\mathbf{X}^\mathsf{T}(\mathbf{I} \otimes \mathbf{\Sigma})^{-1}\mathbf{y} + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right),\ \mathbf{\Sigma}^\star\right)$$

$$\text{where } \ \mathbf{\Sigma}^\star := \left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\mathsf{T}(\mathbf{I} \otimes \mathbf{\Sigma})^{-1}\mathbf{X}\right)^{-1} \in \mathbb{R}^{k,k}$$

We here have:

$$\begin{aligned} \mathbf{\Sigma}^\star &= \left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\mathsf{T}(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}})^{-1}\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{\Sigma}_0^{-1} + (\mathbf{X}_1^\mathsf{T}, \ldots, \mathbf{X}_T^\mathsf{T})(\mathbf{I}^{-1} \otimes \mathbf{\Sigma}_{\mathcal{G}}^{-1}) \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix}\right)^{-1} \\ &= \left(\mathbf{\Sigma}_0^{-1} + \sum_{t=1}^{T} \mathbf{X}_t^\mathsf{T}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\right)^{-1} \end{aligned}$$

and correspondingly:

$$\mathbf{X}^\mathsf{T}(\mathbf{I} \otimes \mathbf{\Sigma}_{\mathcal{G}})^{-1}\mathbf{y} = \sum_{t=1}^{T} \mathbf{X}_t^\mathsf{T}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{y}_t$$

We thus have the posterior distribution:

$$\boldsymbol{\beta}|(\mathbf{\Sigma}_{\mathcal{G}}, \mathbf{y}) \sim \mathcal{N}_k\left(\mathbf{\Sigma}^\star\left(\sum_{t=1}^{T} \mathbf{X}_t^\mathsf{T}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{y}_t + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right),\ \mathbf{\Sigma}^\star\right)$$

where

$$\mathbf{\Sigma}^\star = \left(\mathbf{\Sigma}_0^{-1} + \sum_{t=1}^{T} \mathbf{X}_t^\mathsf{T}\mathbf{\Sigma}_{\mathcal{G}}^{-1}\mathbf{X}_t\right)^{-1}$$

S10 - Additional convergence figures

In this section **S10** we provide exemplary trace plot convergence diagnostics for the mTOR and the ANDRO data. On both data sets we run $H = 10$ independent RJMCMC simulations with $V = 100k$ iterations each. Each RJMCMC simulation yields a trajectory of sampled states:

$$(\mathcal{G}^{(r)}, \mathcal{D}^{(r)}, \boldsymbol{\Sigma}^{(r)}_{\mathcal{G}^{(r)}}, \boldsymbol{\beta}^{(r)}_{\mathcal{D}^{(r)}})_{r=1,\ldots,100k}$$

and we monitor three global quantities, namely

- the 'model posterior score'

$$Q_1^{(r)} := p(\mathbf{Y}|\mathcal{G}^{(r)}, \boldsymbol{\beta}^{(r)}_{\mathcal{D}}) \cdot p(\boldsymbol{\beta}^{(r)}_{\mathcal{D}}) \qquad (r = 1, \ldots, 100k)$$

  which is proportional to the posterior probability of the r-th sampled state.[4]

- the number of contemporaneous (error-error) interactions, symbolically

$$Q_2^{(r)} := \sum_{i=1}^{S} \sum_{j=1}^{S} \mathcal{G}^{(r)}_{i,j} \qquad (r = 1, \ldots, 100k)$$

- the number of dynamic (covariate-response) interactions

$$Q_3^{(r)} := \sum_{i=1}^{N} \sum_{j=1}^{S} \mathcal{D}^{(r)}_{i,j} \qquad (r = 1, \ldots, 100k)$$

For the three quantities the $H = 10$ overlaid trace plots are shown in Figure 1 (mTOR) and Figure 2 (ANDRO). From both figures it can be seen that the trace plots overlay smoothly already after 1k iterations. That is, long before the end of the burn-in phase (of 50k iterations) the trace plots of $H = 10$ independent replicates show oscillations around the same center values and cannot be distinguished anymore. We conclude that none of the trace plots indicates a lack of convergence.
In addition we also considered a few explanatory trace plots of individual model parameters and interaction presence indicators (results not shown). That is, we monitored selected real-valued elements of the model parameter vector $\boldsymbol{\beta}^{(r)}_{\mathcal{D}}$ and selected elements of the matrix $\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}}$, and we monitored selected binary indicator elements of the covariate indicator matrix $\mathcal{D}^{(r)}$ and the DAG $\mathcal{G}^{(r)}$. Like for the three global quantities, $Q_i^{(r)}$ $(i = 1, 2, 3)$ we did not find any evidence for a potential lack of convergence. Potential scale reduction factor based convergence diagnostics are provided in Section 5.1 of the main paper; cf. Figure 1 of the main paper.

[4] Alternative 'model posterior scores' can be defined by marginalizing over $\boldsymbol{\beta}^{(r)}_{\mathcal{D}}$ rather than $\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}}$:

$$Q_{1b}^{(r)} := p(\mathbf{Y}|\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}}, \mathcal{D}^{(r)}) \cdot p(\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}})$$

or by avoiding both marginalizations

$$Q_{1c}^{(r)} := p(\mathbf{Y}|\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}}, \boldsymbol{\beta}^{(r)}_{\mathcal{D}}) \cdot p(\boldsymbol{\beta}^{(r)}_{\mathcal{D}}) \cdot p(\boldsymbol{\Sigma}_{\mathcal{G}^{(r)}})$$

For all three 'model posterior scores' we observed smoothly overlaying trace plots without any trends.

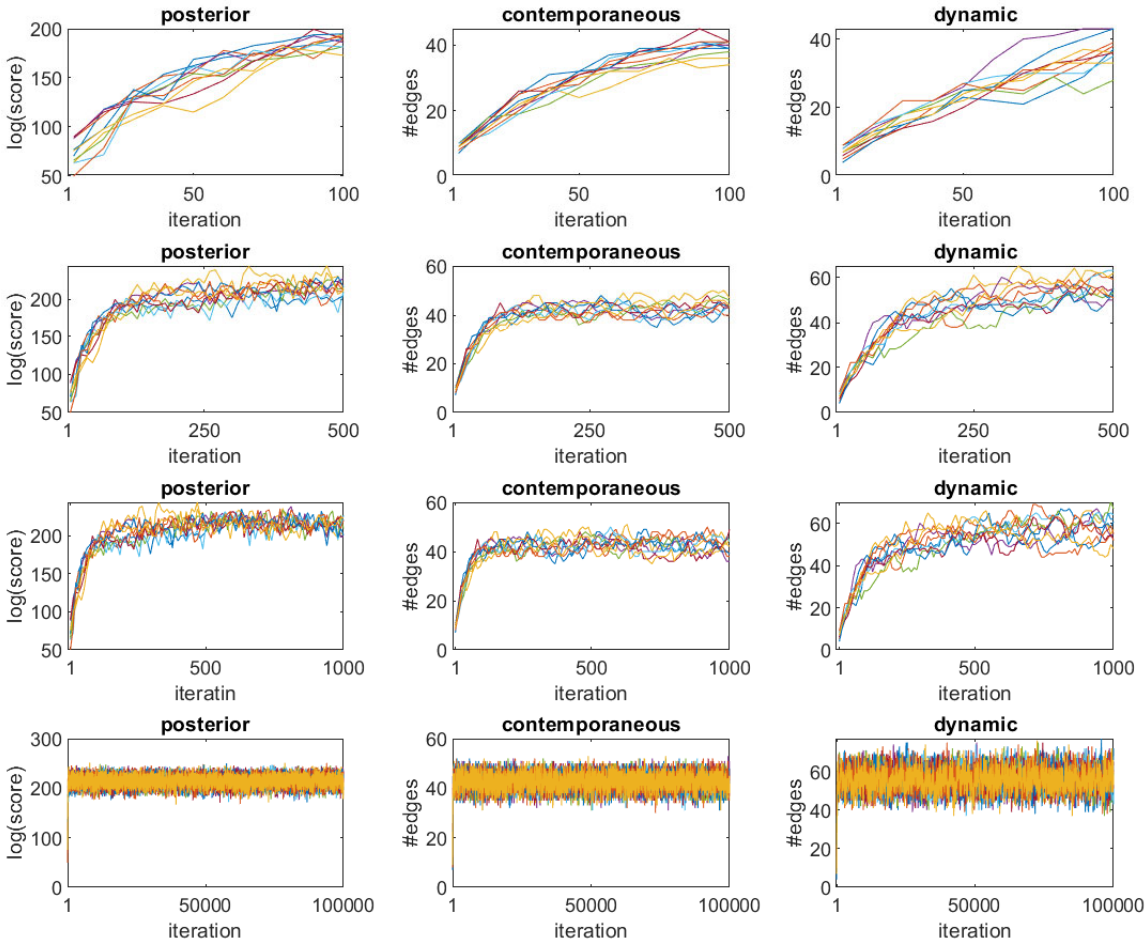

Figure 1: **Additional figure 1 - Trace plot diagnostics for mTOR data**. Each panel shows overlaid trace plots, obtained from $H = 10$ independent MCMC simulations of length 100k on the mTOR data. We thinned out each trajectory by a factor of 10. The trace plots monitor the logarithmic 'model posterior scores' $\log(Q_1^{(r)})$ (left column), the number of static edges $Q_2^{(r)}$ (center column) and the number of dynamic edges $Q_3^{(r)}$ (right column). The four rows refer to four different sections of the trajectory, focusing only on the results of the first 0.1k, 0.5k, 1k iterations, and finally covering the whole trajectory of $V = 100k$ iterations (bottom row).

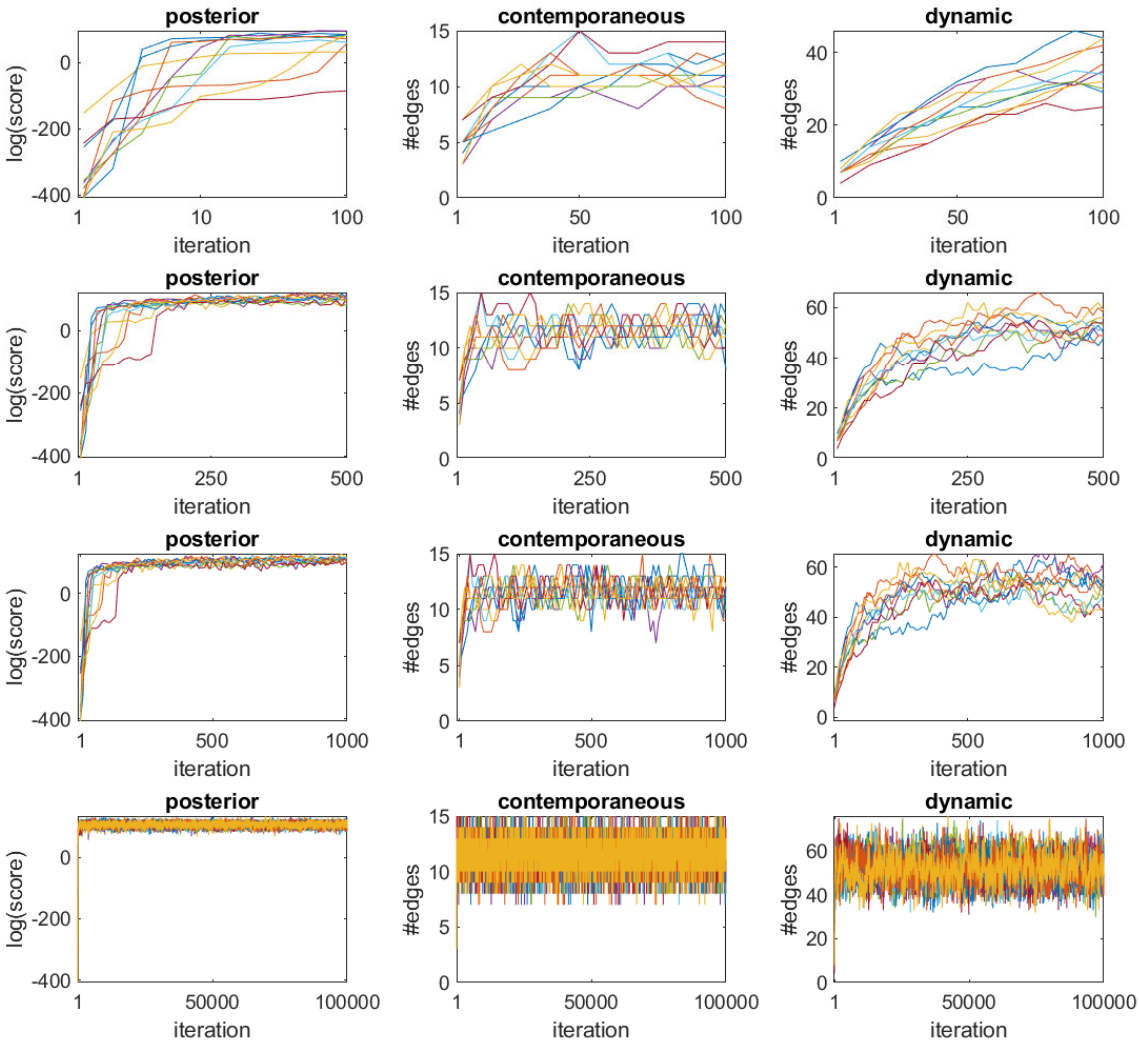

Figure 2: **Additional figure 2 - Trace plot diagnostics for ANDRO data**. Each panel shows overlaid trace plots, obtained from $H = 10$ independent MCMC simulations of length 100k on the ANDRO data. We thinned out each trajectory by a factor of 10. The trace plots monitor the logarithmic 'model posterior scores' $\log(Q_1^{(r)})$ (left column), the number of static edges $Q_2^{(r)}$ (center column) and the number of dynamic edges $Q_3^{(r)}$ (right column). The four rows refer to four different sections of the trajectory, focusing only on the results of the first 0.1k, 0.5k, 1k iterations, and finally covering the whole trajectory of $V = 100k$ iterations (bottom row).

S11 - Additional figures for synthetic data

In this section **S11** of the supplementary material we provide the results of two additional studies on the synthetic data (cf. Sect. 5.2.1 of the main paper). In Figs. 3-5 we study the effect of alternative hyperparameter settings with stronger penalties for the regression parameter sizes ($\alpha \in \{12, 24, 36, 48\}$). From Fig. 3 (predictive probabilities) and Fig. 5 (AUCs) it can be seen that the trends from Fig. 3 of the main paper are conserved. In both criteria (predictive probabilities and AUCs) the differences stay significantly in favour of the $\mathcal{M}_{1,1}$ model; only the relative differences decrease as α increases. Fig. 4 shows for each model how the predictive probabilities vary with α. The predictive probabilities of the models $\mathcal{M}_{1,1}$ and $\mathcal{M}_{1,0}$ decrease as the prior penalty gets larger, while the predictive probabilities of the models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{0,1}$ increase from $\alpha = 12$ to $\alpha = 24$ and then run into a plateau. These trends show the positive effect of model selection (averaging) over 'full' model analyses. The over-complex $\mathcal{M}_{0,0}$ model includes all possible interactions and, henceforth, requires sparse/restrictive parameter priors to keep the spurious interactions under control. The proposed $\mathcal{M}_{1,1}$ model infers the model (sets of interactions) and this way gets rid of spurious interactions. Unlike the $\mathcal{M}_{0,0}$ model, the new $\mathcal{M}_{1,1}$ model thus does not require sparse/restrictive parameter priors and performs best for the least restrictive parameter priors.

In Fig. 6 we compare the proposed $\mathcal{M}_{1,1}$ model with the traditional MBLR model. For the traditional MBLR model (cf. Sect. 2.1 of the main paper) we have the priors

$$\begin{aligned} \boldsymbol{\Sigma} &\sim \mathcal{W}_S^{-1}(\mathbf{V}, \alpha) \\ \text{vec}(\mathbf{B}) &\sim \mathcal{N}_{k \cdot s}(\text{vec}(\mathbf{B}_0), \boldsymbol{\Sigma} \otimes \Delta_0^{-1}) \end{aligned}$$

To be consistent with the $\mathcal{M}_{i,j}$ models, we set $\mathbf{V} = \frac{1}{\alpha - S - 1} \cdot \mathbf{I}$, $\alpha = 12$, and $\text{vec}(\mathbf{B}_0) = \mathbf{0}$. Moreover, we set $\Delta_0^{-1} := \delta \mathbf{I}$ and we vary the tuning hyperparameter δ over a large range of values ($\delta = \left(\frac{1}{3}\right)^w$ with $w = -6, \ldots, 6$).
From the left panel of Fig. 6 it can be seen that the predictive probabilities of the traditional MBLR model (TRAD) are consistently lower than those of the $\mathcal{M}_{1,1}$ model and get substantially worse for large δ values. The center panel shows that the differences in the log predictive probabilities are significantly in favour of the $\mathcal{M}_{1,1}$ model. The right panel of Fig. 6 compares the AUC values of the models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{1,1}$ with the AUCs of the best performing traditional MBLR model ($\delta = 1/81$). The AUCs of the traditional MBLR model are lower than those of the $\mathcal{M}_{0,0}$ model, and hence clearly inferior to the AUCs of the $\mathcal{M}_{1,1}$ model. The most noteworthy finding is that the traditional MBLR model performs rather bad in learning the error-error interactions. A possible explanation for this shortcoming is that the traditional MBLR re-employs the error covariance matrix $\boldsymbol{\Sigma}$ in the regression parameter prior (see equations above). This way the regression parameters interfere with the error covariance matrix $\boldsymbol{\Sigma}$, what might introduce a bias and explain the lower accuracy in learning the error-error interactions.

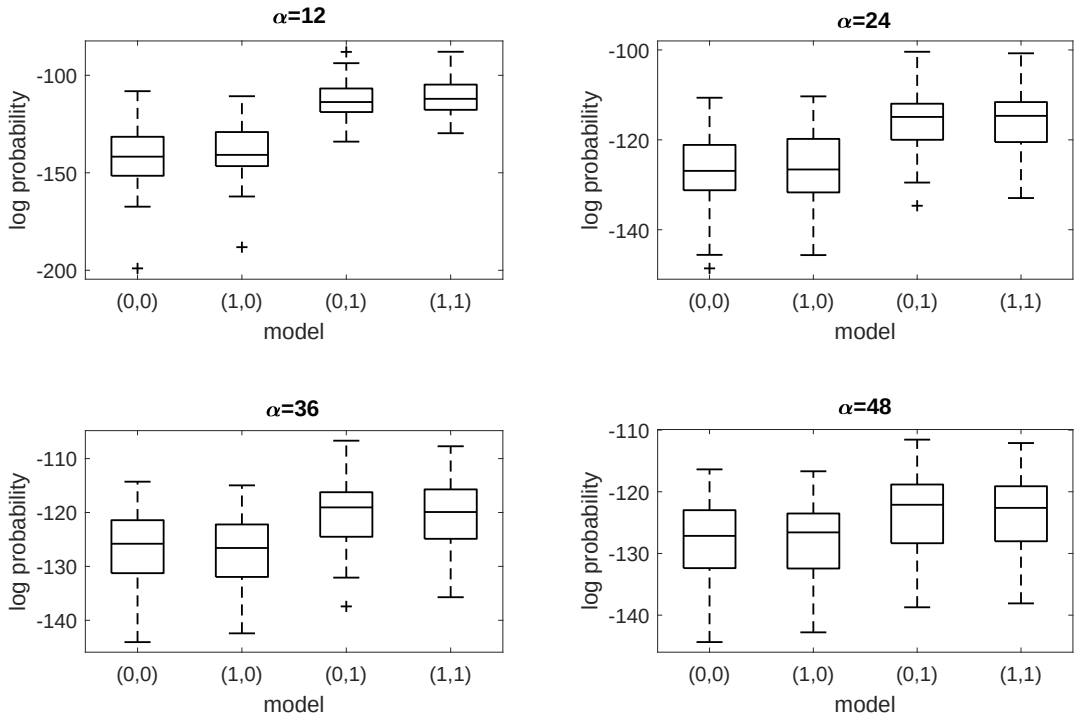

Figure 3: **Additional figure 3 (synthetic data).** The four panels show box plots of the log predictive probabilities for four hyperparameter values $\alpha \in \{12, 24, 36, 48\}$. Each panel shows the four model-specific boxplots. We note that the top left panel refers to the top left panel of Fig. 1 of the main paper, and we refer to its caption for more technical details.

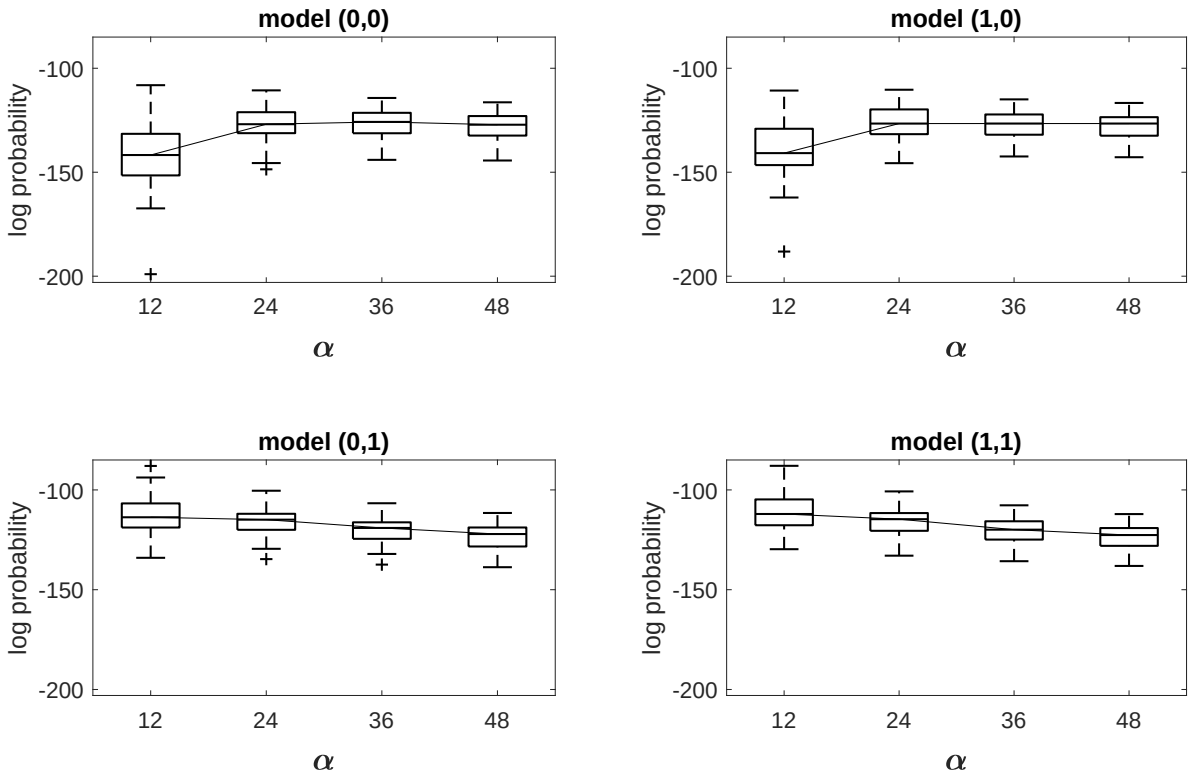

Figure 4: **Additional figure 4 (synthetic data).** For each of the fours models under comparison $\mathcal{M}_{i,j}$ $(i, j \in \{0, 1\})$ there is a panel showing how the log predictive probabilities vary with the hyperparameter $\alpha \in \{12, 24, 36, 48\}$. See caption of Fig. 1 of the main paper for more details.

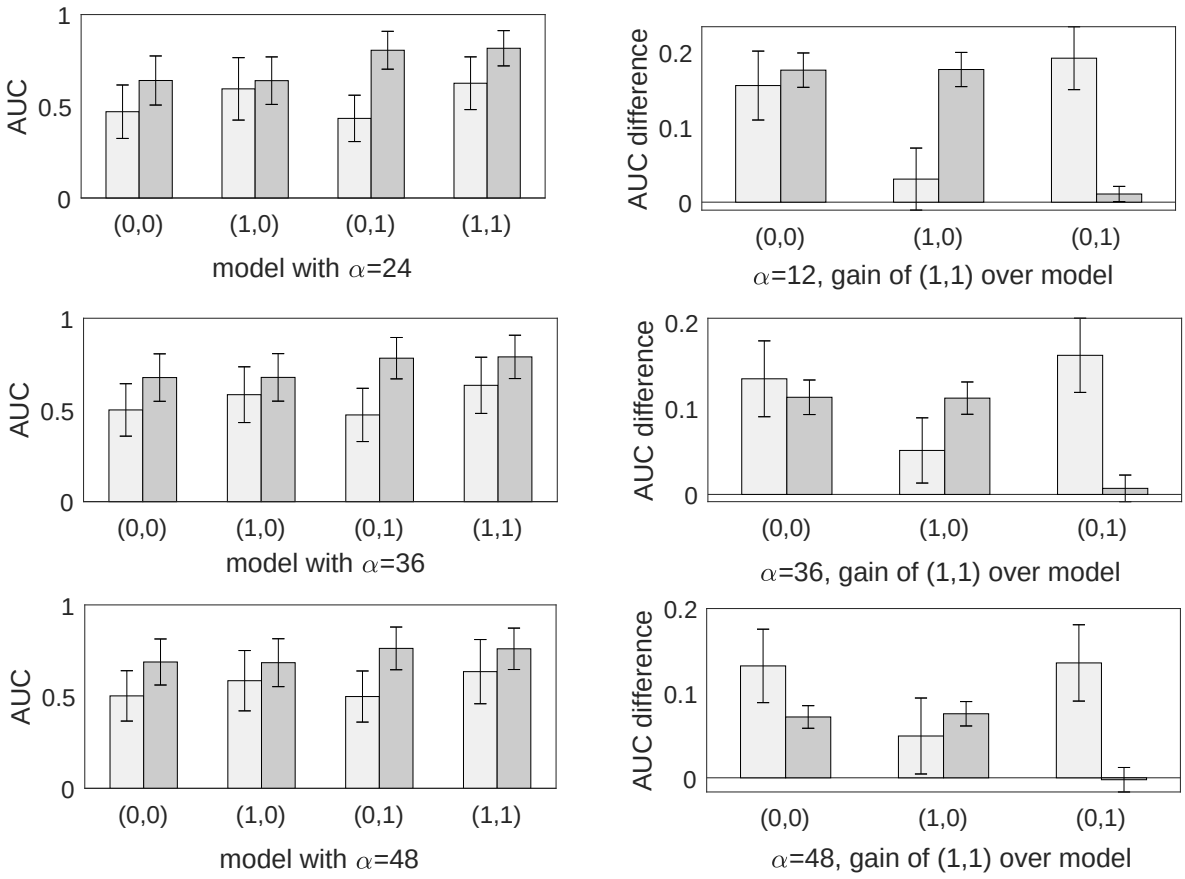

Figure 5: **Additional figure 5 (synthetic data).** Areas under the precision recall curves (AUCs) for three alternate hyperparameter settings $\alpha \in \{24, 36, 48\}$. The panels in the left column show the model-specific AUCs, and the panels in the right column show the relative gains of model $\mathcal{M}_{1,1}$. We computed separate AUCs for the error-error ($e_i \to e_j$, light grey) and the covariate-response ($z_i \to y_s$, dark grey) interactions. The AUC results for $\alpha = 12$ are shown in Fig. 1 of the main paper and we refer to its caption for more technical details.

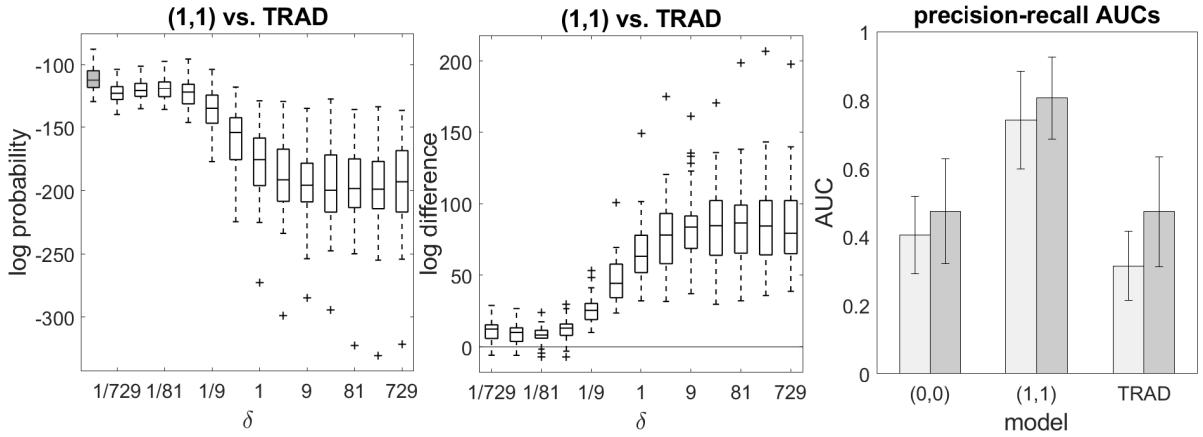

Figure 6: **Additional figure 6 (synthetic data).** We applied the traditional MBLR model (TRAD) using the priors $\mathbf{\Sigma} \sim \mathcal{W}_S^{-1}(\mathbf{I}, 12)$ and $\text{vec}(\mathbf{B}) \sim \mathcal{N}_{k \cdot s}(\mathbf{0}, \mathbf{\Sigma} \otimes \delta\mathbf{I})$ for varying hyperparameters $\delta = \left(\frac{1}{3}\right)^w$ with $w = -6, \ldots, 6$. The left panel shows boxplots of the log predictive probabilities and begins with the boxplot of the $\mathcal{M}_{1,1}$ model (in grey) followed by 13 boxplots for the traditional model (in white). The center panel shows the relative differences in the log predictive probabilities in favour of the $\mathcal{M}_{1,1}$ model. The right panel compares the AUC values of the best TRAD model ($\delta = 1/81$) with the AUC values of the two models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{1,1}$. The figure shows separate AUCs for the error-error (light grey) and the covariate-response (dark grey) interactions. Interestingly, the error-error AUC of the best traditional model (TRAD) is lower then the error-error AUC value of the akin $\mathcal{M}_{0,0}$ baseline model.

S12 - Additional figures for mTOR/ANDRO data

In this section **S12** of the supplementary material we provide more results for the mTOR and for the ANDRO data (cf. Sect. 5.2.2 of the main paper). Instead of comparing with the Bayesian network types (dynamic $\mathcal{M}_D$ and static $\mathcal{M}_S$), we here cross-compare the performances of the four MBLR model variants $\mathcal{M}_{i,j}$ $(i, j \in \{0, 1\})$. For both data sets mTOR (see Fig. 7) and ANDRO (see Fig. 8) we observe very similar results. The proposed $\mathcal{M}_{1,1}$ outperforms the baseline model $\mathcal{M}_{0,0}$ and the 'in between' model $\mathcal{M}_{1,0}$, but it does not improve over the $\mathcal{M}_{0,1}$ model. This suggests that working with a full (over-complex) error DAG has no negative effects on the predictive probabilities, while working with a full (over-complex) covariate matrix yields significantly worse performances. This finding is consistent with the result for the synthetic data, where we also found that the $\mathcal{M}_{0,1}$ model performed second best in terms of predictive probabilities. A possible explanation are the mismatches in the numbers of possible covariate-response and error-error interactions. For a data set with N covariates and S responses, there are $\mathcal{I}_1 = N \cdot S$ possible covariate-response interactions and $\mathcal{I}_2 = S \cdot (S+1)/2$ possible error-error interactions. Table 1 lists the numbers of possible interactions for all data sets, and it can be seen that all ratios are clearly in favour of the covariate-response interactions. Therefore using a full covariate matrix potentially introduces more spurious interactions than using a full error DAG.

Table 1: **Overview: Possible interactions per data set type**. For each data set type the table lists the numbers of covariates N, the numbers of responses S, the number of possible covariate-response interactions ($\mathcal{I}_1$), the number of possible error-error interactions ($\mathcal{I}_2$), and the ratio $\mathcal{I}_1/\mathcal{I}_2$.

data	N	S	$\mathcal{I}_1$	$\mathcal{I}_2$	Ratio
Synthetic	20	10	200	45	4,44
mTOR	11	11	121	55	2.20
ANDRO	30	6	180	15	12.00
OES10	20	20	400	190	2.11

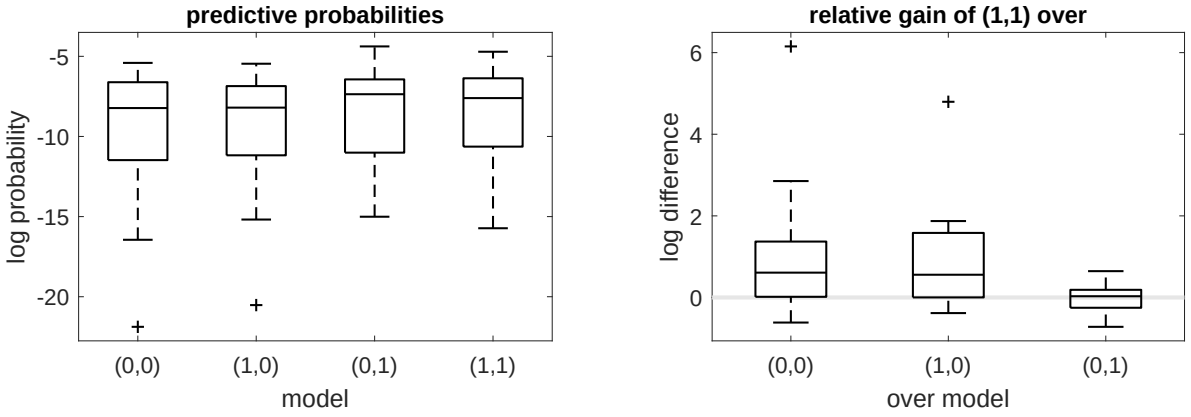

Figure 7: **Additional figure 7 (mTOR data).** We applied leave-one-out-cross-validation to compute a predictive probability for each of the 18 data points. The left panel shows box plots of the 18 predictive probabilities achieved with the four models $\mathcal{M}_{i,j}$ $(i, j \in \{0, 1\})$. The right panel shows the relative gains of model $\mathcal{M}_{1,1}$ over the three competitors. The three differences (in median) lead to the Wilcoxon signed-rank p-values: 0.004, 0.002 and 1.000, showing that the performances of $\mathcal{M}_{1,1}$ and $\mathcal{M}_{0,1}$ do not significantly differ.

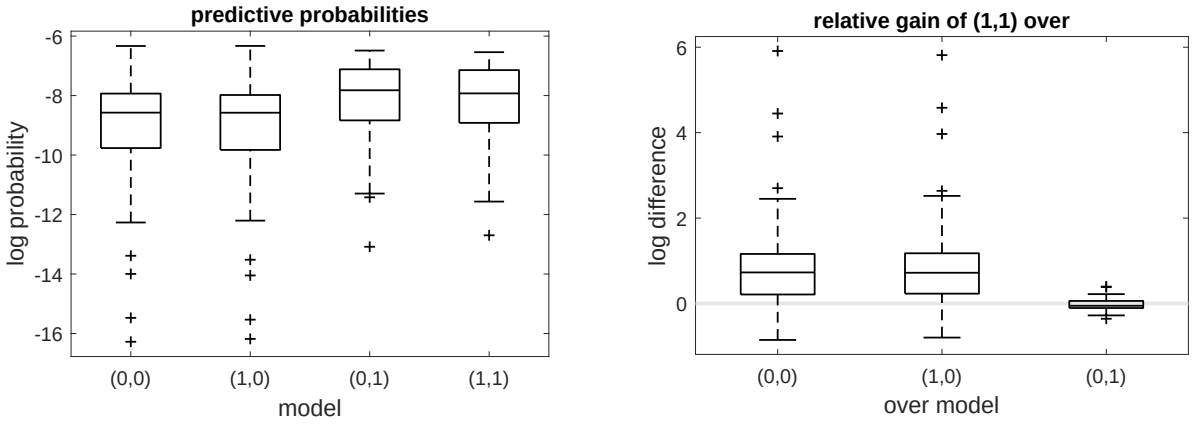

Figure 8: **Additional figure 8 (ANDRO data).** We applied leave-one-out-cross-validation strategy to compute a predictive probability for each of the 49 data points. The left panel shows box plots of the 49 predictive probabilities achieved with the four models $\mathcal{M}_{i,j}$ $(i, j \in \{0, 1\})$. The right panel shows the relative gains of model $\mathcal{M}_{1,1}$ over the three competitors. The three differences (in median) lead to the Wilcoxon signed-rank p-values: $9.0 \cdot 10^{-8}$, $6.1 \cdot 10^{-8}$ and 0.611, showing that the performances of $\mathcal{M}_{1,1}$ and $\mathcal{M}_{0,1}$ do not significantly differ.

S13 - Additional figures for OES 2010 data

In this section **S13** of the supplementary material we provide three additional figures for the OES 2010 data (cf. Sect. 5.2.3 of the main paper). Fig. 9 shows boxplots of the model-specific log predictive probabilities for the six different sample sizes $T \in \{10, 20, \ldots, 320\}$. From those log predictive probabilities we computed the relative differences shown in Fig. 6 of the main paper. For the proposed $\mathcal{M}_{1,1}$ model, Fig. 10 monitors how the distribution of the log predictive probabilities changes as the sample size increases. As the slopes of the lines that connect the boxplot medians seems to decrease, we conclude that the improvement rate gets weaker as the sample size increases. That is, doubling a small sample size brings more improvement in the predictive probabilities than doubling a large sample size. This indicates that the predictive probabilities run into a plateau, where increasing the sample size does not lead to noteworthy improvements anymore. Fig. 11 presents the same boxplots like Fig. 6 of the main paper but in another arrangement. Instead of six panels for the sample sizes, each containing three pairwise model comparisons, Fig. 11 has a panel for each of the three model comparisons and monitors how the superiority of the proposed $\mathcal{M}_{1,1}$ model diminishes as the sample size T increases.

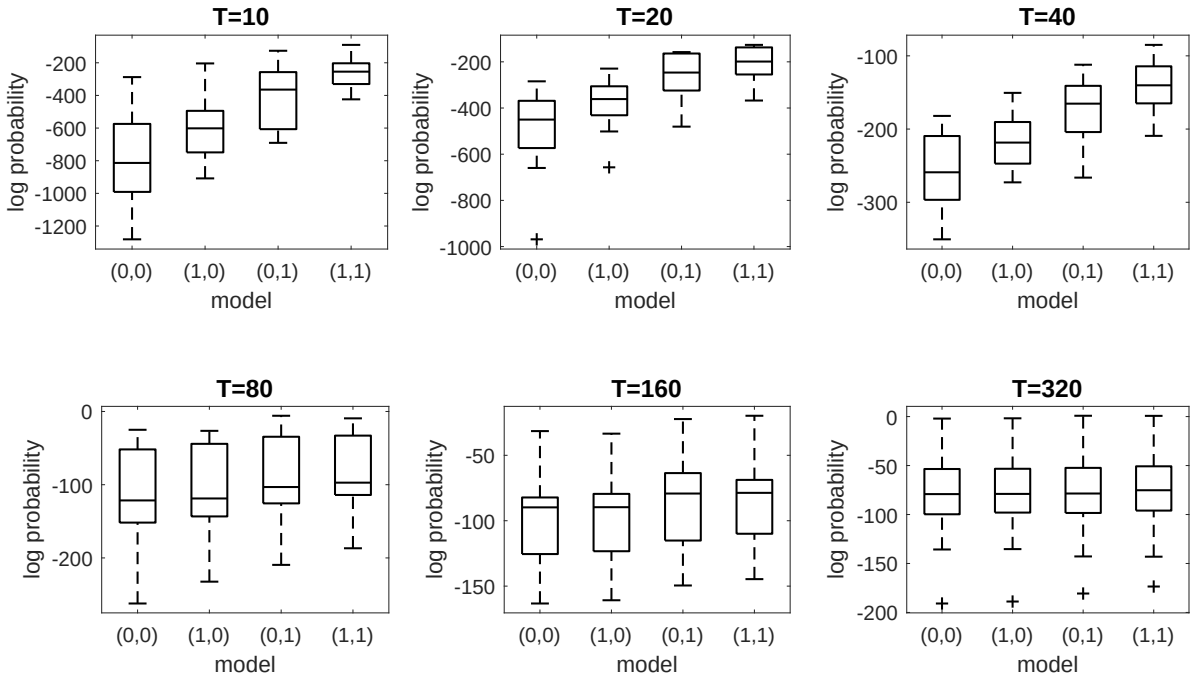

Figure 9: **Additional figure 9 for OES 2010 data.** Each panel refers to a sample size T and shows box plots of the log predictive probabilities for the four model variants $\mathcal{M}_{i,j}$ $(i, j \in \{0, 1\})$. Each boxplot summarises 10 predictive probabilities obtained for 10 independent training data sets of size T and validation data sets of size $\tilde{T} = 10$. From the log predictive probabilities (shown here) we computed the relative differences, whose boxplots are shown in Fig. 6 of the main paper.

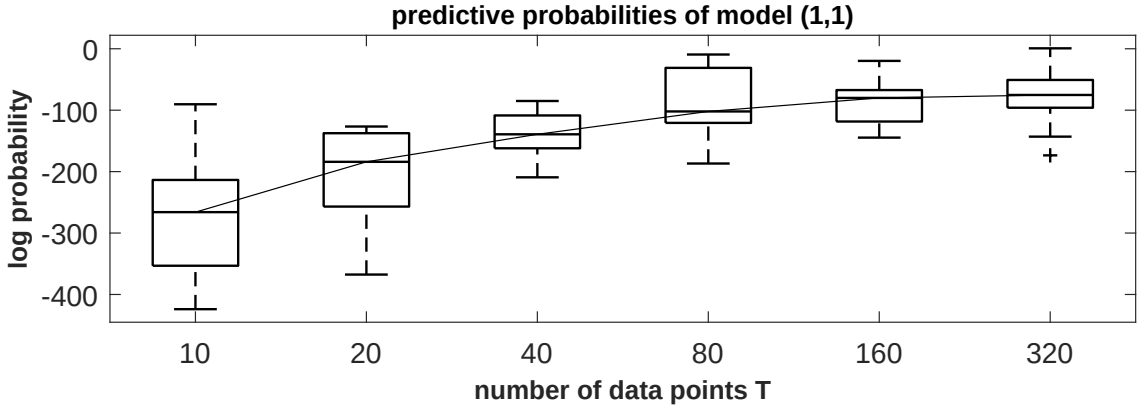

Figure 10: **Additional figure 10 for OES 2010 data**. The figure shows how the log predictive probabilities of the proposed $\mathcal{M}_{1,1}$ model increase with the sample size T of the training data sets. For each T we used validation data sets with the sample size $\tilde{T} = 10$ to compute the predictive probabilities. For the three other model variants $\mathcal{M}_{i,j}$ we observed similar trends.

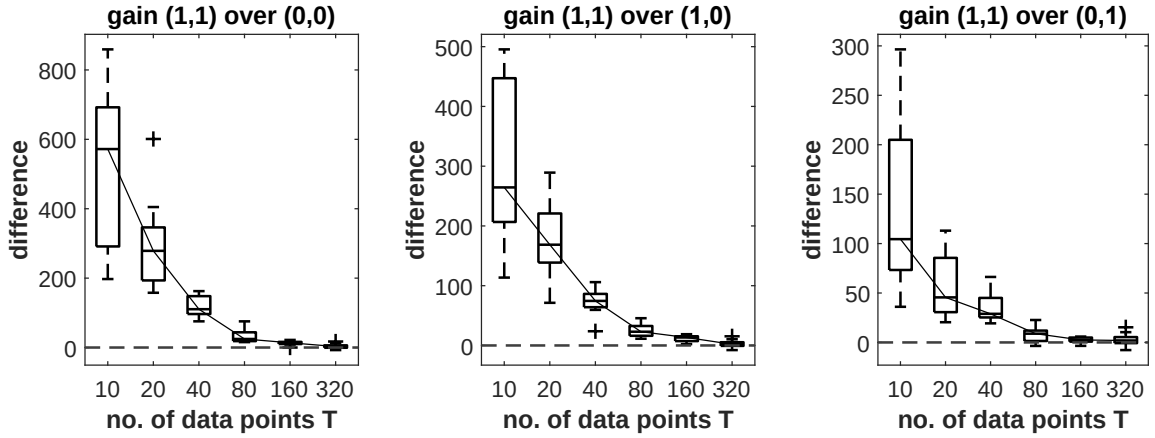

Figure 11: **Additional figure 11 for OES 2010 data**. Re-arrangement of the box plots from Fig. 6 of the main paper. Unlike Fig. 6 of the main paper, this figure has a panel for each pairwise model comparison $\mathcal{M}_{1,1}$ vs. $\mathcal{M}_{0,0}$ (left), $\mathcal{M}_{1,1}$ vs. $\mathcal{M}_{1,0}$ (center), and $\mathcal{M}_{1,1}$ vs. $\mathcal{M}_{0,1}$ (right). Each panel shows box plots of the relative log differences and monitors how the superiority of the proposed $\mathcal{M}_{1,1}$ model diminishes as the sample size T increases.

S14 - Strategies to reduce the computational costs

The measured computational times in Table 2 of the main paper suggest that our Matlab implementation of the RJMCMC algorithm (see Sect 2.4 of the main paper) does not scale well. In this section **S14** of the supplement we therefore briefly list possible methodological approaches and approximations that could be considered for reducing the computational costs and improving the scalability.
Apart from implementing the code in a more efficient programming language (e.g. C++) or parallelizing the code along the lines of Hans et al. (2007), one could consider:

- **Informative RJMCMC initialisation:**
  To reduce the number of required RJMCMC iterations, one could consider initialising the RJMCMC sampler with an informative model. A good starting model could, for example, be extracted from the results of the computationally cheaper traditional MBLR model.

- **Hard fan-in constraints:**
  One could impose restrictions on the maximal sizes of the covariate sets and/or parent sets in the error DAG. Such fan-in restrictions are widely applied in static Bayesian networks to restrict the model configuration space, so as to reduce the computational costs. For examples we refer to Friedman and Koller (2003) and Grzegorczyk and Husmeier (2008).

- **More restrictive model priors:**
  Instead of imposing hard constraints (fan-in's), one could replace the uniform model prior by a stringent prior that penalises the numbers of covariate-response and error-error interactions, so as to keep the model complexity (= the number of interactions) low.

- **Interaction pre-filtering:**
  Via univariate data pre-analyses one could try to identify the most promising interactions and rule out all other interactions in a subsequent multivariate analysis with the $\mathcal{M}_{1,1}$ model. Covariate-response interactions could be filtered out via univariate linear regression analyses. Error-error interactions could be filtered out by pre-computing the local scores $p(\mathbf{E}^{y_i|y_j})$ of the $S(S-1)$ possible response-response interactions (cf. Eq. (19) of the main paper), and ruling out the lowest scoring response-response (error-error) interactions; see, e.g., Friedman and Koller (2003) for an example algorithm.

- **Working with 'in-between' models:**
  Our empirical results suggest that the $\mathcal{M}_{1,0}$ model (with full error DAG) does not perform much worse than the $\mathcal{M}_{1,1}$ model. In particular for large sample sizes T one could thus resort to the $\mathcal{M}_{1,0}$ model, which does not require the error DAG to be learned. In the same vein, to substantially reduce the configuration space of possible models, one could adopt the idea of Yang et al. (2020) and impose the restriction that all responses share the same covariate set rather than allowing for response-specific covariate sets.

- **Using distributed Bayes:**
  A long-term perspective might be to divide a large data set into parts and to resort to distributed Bayes techniques, so as to analyse smaller data subsets in parallel and to approximate the overall result from the individual results. See, for example, the recently proposed approaches in Song et al. (2020) and Buchholz et al. (2019). However, in the context of RJMCMC sampling distributed Bayes methods have not been well studied yet; see, e.g., the discussion in Sect. 3.5 of Buchholz et al. (2019).

References

Bishop CM (1995) Neural Networks for Pattern Recognition. Oxford University Press, New York, iSBN 0-19-853864-2

Buchholz A, Ahfock D, Richardson S (2019) Distributed computation for marginal likelihood based model choice. arxiv191004672

Friedman N, Koller D (2003) Being Bayesian about network structure. Machine Learning 50:95–126

Geiger D, Heckerman D (1994) Learning Gaussian networks. In: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA., pp 235–243

Geiger D, Heckerman D (2002) Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics 30(5):1412–1440

Grzegorczyk M, Husmeier D (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning 71(2-3):265–305

Hans C, Dobra A, West M (2007) Shotgun stochastic search in regression with many predictors. Journal of the American Statistical Association 102(478):507–516

Shachter RD, Kenley R (1989) Gaussian influence diagrams. Management Science 35:527–550

Song Q, Sun Y, Ye M, Liang F (2020) Extended stochastic gradient Markov chain Monte Carlo for large-scale Bayesian variable selection. Biometrika 107(4):997–1004

Yang D, Goh G, Wang H (2020) A fully Bayesian approach to sparse reduced-rank multivariate regression. Statistical Modelling DOI https://doi.org/10.1177/1471082X20948697, onlineFirst